

# *Bayesian estimates of autocorrelations in single-case designs*

**William R. Shadish, David M. Rindskopf, Larry V. Hedges & Kristyann J. Sullivan**

**Behavior Research Methods**

e-ISSN 1554-3528

Behav Res  
DOI 10.3758/s13428-012-0282-1

**Behavior  
Research  
Methods**

VOLUME 44, NUMBER 4 ■ DECEMBER 2012



# BRM

**EDITOR**

Gregory Francis, *Purdue University*

**ASSOCIATE EDITORS**

Ira H. Bernstein, *University of Texas Southwest Medical Center*  
Mark W. Greenlee, *University of Regensburg*  
Kim Vu, *California State University Long Beach*

A PSYCHONOMIC SOCIETY PUBLICATION  
[www.psychonomic.org](http://www.psychonomic.org)  
ISSN 1554-3528



 Springer

**Your article is protected by copyright and all rights are held exclusively by Psychonomic Society, Inc.. This e-offprint is for personal use only and shall not be self-archived in electronic repositories. If you wish to self-archive your work, please use the accepted author's version for posting to your own website or your institution's repository. You may further deposit the accepted author's version on a funder's repository at a funder's request, provided it is not made publicly available until 12 months after publication.**

# Bayesian estimates of autocorrelations in single-case designs

William R. Shadish · David M. Rindskopf ·  
Larry V. Hedges · Kristynn J. Sullivan

© Psychonomic Society, Inc. 2012

**Abstract** Researchers in the single-case design tradition have debated the size and importance of the observed autocorrelations in those designs. All of the past estimates of the autocorrelation in that literature have taken the observed autocorrelation estimates as the data to be used in the debate. However, estimates of the autocorrelation are subject to great sampling error when the design has a small number of time points, as is typically the situation in single-case designs. Thus, a given observed autocorrelation may greatly over- or underestimate the corresponding population parameter. This article presents Bayesian estimates of the autocorrelation that greatly reduce the role of sampling error, as compared to past estimators. Simpler empirical Bayes estimates are presented first, in order to illustrate the fundamental notions of autocorrelation sampling error and shrinkage, followed by fully Bayesian estimates, and the difference between the two is explained. Scripts to do the analyses are available as supplemental materials. The analyses are illustrated using two examples from the single-case design literature. Bayesian estimation warrants wider use, not only in debates about the size of autocorrelations, but also in statistical methods that require an independent estimate of the autocorrelation to analyze the data.

**Electronic supplementary material** The online version of this article (doi:10.3758/s13428-012-0282-1) contains supplementary material, which is available to authorized users.

W. R. Shadish (✉) · K. J. Sullivan  
School of Social Sciences, Humanities and Arts,  
University of California, Merced,  
5200 North Lake Rd,  
Merced, CA 95343, USA  
e-mail: wshadish@ucmerced.edu

L. V. Hedges  
Institute for Policy Research, Northwestern University,  
Evanston, IL, USA

D. M. Rindskopf  
Graduate Center, City University of New York,  
New York, NY, USA

**Keywords** Bayesian estimation · Autocorrelation · Single-case designs

Single-case designs (SCDs) are widely used to assess the impact of interventions in fields as diverse as medicine, developmental disabilities, education, and behavioral disorders (Gabler, Duan, Vohra, & Kravitz, 2011; Shadish & Sullivan, 2011). They generally take the form of a short, interrupted time series design in which an intervention is applied to a single case that is observed for a particular outcome on many occasions over time. Historically, researchers who use SCDs have relied extensively on visual analysis to assess whether the intervention has had the desired impact on the outcome. In the last two decades, however, efforts to develop and apply statistical analyses to SCD data have increased greatly (Kratochwill & Levin, 2010; Maggin et al., 2011; Parker, Vannest, & Davis, 2011; Parker, Vannest, Davis, & Sauber, 2011; Shadish & Rindskopf, 2007; Shadish, Rindskopf, & Hedges, 2008). One common attribute of these diverse statistical efforts is the need to take into account the serial dependency among errors of observation within a case over time, commonly referred to as *autocorrelation*. This dependence violates the assumption of independence of errors that is shared by nearly all parametric and nonparametric statistics, and it can result in biased descriptive and inferential statistics, the latter typically resulting in an inflated Type I error rate.

SCD researchers have long been aware of this problem. Over 30 years ago, for example, Jones, Weinrott, and Vaught (1978) explored the effects of serial dependency on both visual and statistical inference. Since then, many researchers have explored methods for measuring autocorrelation, surveyed the observed size of autocorrelations in the SCD literature, gauged the potential impact of autocorrelation on effect estimation, and suggested ways to minimize that impact (e.g., Huitema & McKean, 1994; Manolov & Solanas, 2008; McKnight, McKean, & Huitema, 2000). Plaguing all of these efforts, however, have been two special

problems in measuring autocorrelation, problems that are particularly prevalent in SCDs. Both problems stem from the fact that SCDs tend to be short; that is, they have comparatively few observations within a case over time, as compared to the traditional time series literature, in which 50–100 observations over time are common. First is the problem of bias—that estimates of autocorrelation tend to be negatively biased in short time series. This problem has received the most attention in the literature. For instance, Huitema and his colleagues have worked to develop estimators that may be less sensitive to this bias (Huitema & McKean, 1994; McKnight et al., 2000). Second is the problem of precision—that autocorrelations from any given SCD are subject to substantial fluctuation due to sampling error, so that the variance of observed estimates around the point estimate is very high. The present article addresses the issue of precision and shows that Bayesian autocorrelation estimators may improve our understanding of the precision of any of the extant autocorrelation estimators. Unless explicitly noted otherwise, we use the word *autocorrelation* to refer to the lag-1 correlation among residuals from observations, frequently obtained empirically by a regression of those observations on time, treatment, and a term to represent the interaction (Huitema & McKean, 2000).

### More on precision: Population parameters, sample statistics, and sampling error

*Sampling error* is the discrepancy between a population parameter and the sample realizations of that parameter that may be observed in any given data set. The autocorrelation is as subject to sampling error as any other observed statistic, and sampling error leads to a wider range of observed autocorrelations than many researchers may realize. For example, Parker et al. (2005) measured the observed autocorrelation in 77 published SCDs and reported that the middle 80 % of the distribution ranged from about  $r = -.33$  to about  $r = .77$ , depending on how the autocorrelation was measured. They concluded that autocorrelation exists in large amounts and may be quite problematic for statistical analysis. Yet much of this observed range was likely due to sampling error. With  $k$  cases (time series), each with a population autocorrelation, the asymptotic variance of the estimated autocorrelation is approximately

$$v_j = (1 - \rho_j^2) / (t_j - 3), \quad (1)$$

where  $\rho$  is the autocorrelation of the  $j$ th case ( $j = 1 \dots k$ ) and  $t$  is the number of time points in the  $j$ th case (Anderson, 1971). Suppose that the population autocorrelation is actually  $\rho = .222$ . Then a range of  $r = -.33$  to about  $r = .77$  for the middle 80 % of

the observations is the variability we would expect by chance if the set of SCDs averaged 8.15 observations. We used 80 % probability intervals to match what Parker et al. (2005) had done, but it would be more common to use 95 % probability intervals, which we will do in the next few examples. With an SCD that has 20 time points, we would expect that the autocorrelation in the observed data might vary by chance from  $-.26 < r_j < .66$  (the 95 % probability interval). With only ten time points, which is not at all uncommon in the SCD literature, the range would be  $-.53 < r_j < .93$ . Even with as many as 35 time points, which is more than most SCDs have, the range would still be  $-.14 < r_j < .54$ . Clearly the range of autocorrelations described by Parker et al. (2005) might be due substantially to chance.

Hence, the autocorrelation computed for any given SCD is a very imprecise estimate of the population autocorrelation that produced the data—even if it were an unbiased estimate. Fortunately, when autocorrelations are computed from multiple SCDs within a study, we can use that larger set of information to obtain better autocorrelation estimates in two ways. The first is to use random effects meta-analytic methods to estimate the population parameter(s) and then to use the value(s) in the pertinent statistical analyses. We have used that option in other research, not reported here (Hedges, Pustejovsky, & Shadish, 2012), but it has a problem: Unless the variance component measuring the heterogeneity of the observed autocorrelations is zero, using the meta-analytic mean incorrectly implies that all cases within a study have exactly the same autocorrelation, which may not be the case if homogeneity is rejected.

Hence, the present article proposes a second option, using Bayesian statistics to obtain empirical Bayes or fully Bayesian estimates of the autocorrelation. We start by presenting the simpler, empirical Bayes estimates, which show how Bayesian statistics can take advantage of information about the autocorrelation from all of the cases in a study to reduce the role of sampling error in estimates of the autocorrelation for any single case. When the computed estimate of the heterogeneity of the autocorrelation is zero, empirical Bayes estimates shrink the individual estimates for each case to a single common mean, because all of the apparent variability in individual estimates is due to sampling error. When heterogeneity is greater than zero, estimates are not shrunk as much, and if heterogeneity is truly very large, estimates may not be shrunk at all. Empirical Bayes estimates are simple to compute with the SPSS and R code that we provide in the online [supplemental materials](#).

However, empirical Bayes methods assume that the heterogeneity of the autocorrelation is measured perfectly, whether or not it happens to be zero. Fully Bayesian methods treat the heterogeneity not as a single parameter represented by a point estimate, but as itself having a distribution in the population. The result is not a point estimate of the

autocorrelation, but rather a posterior distribution that shows how estimates of the autocorrelation within a study vary, depending on uncertainty about heterogeneity. Although fully Bayesian methods are more difficult to compute for many applied researchers because they require the use of programs such as WinBUGS or the hblm module of SPSS (code for the latter is also available in the online [supplemental materials](#)), the effort is worthwhile, given the power that fully Bayesian methods have to illuminate problems of autocorrelation in SCDs.

### Empirical Bayes estimates

Studies that use SCD methods typically include more than one SCD (Shadish & Sullivan, 2011). If one has observed autocorrelations from both a given SCD within a study and a larger set of SCDs from the same study that are considered exchangeable with that given SCD, the empirical Bayes (EB) estimate is an optimal composite based on data from both sources. The exchangeability assumption implies that the researcher has no prior reason to assume that the autocorrelation for one case within a study should be larger or smaller than that of any other case for reasons other than sampling error or systematic differences due to the measured covariates. This seems plausible in the SCD literature. In most SCD studies, the researcher is investigating the effects of a common treatment on a set of cases selected to be similar in such study-relevant characteristics as diagnosis, age, gender, or disability, and the researcher is measuring those effects on the same outcome over time. However, this assumption may not hold, and sometimes Bayesian estimates can shed light on that occurrence, as we will show in one example.

Here we use the usual Yule–Walker estimate of autocorrelation, uncorrected for bias due to small sample size or the number of parameters used to estimate the residuals on which the autocorrelation is computed. We have not used the bias-corrected versions of the autocorrelation and its variance, for two reasons. First, for purposes of demonstrating a Bayesian approach to this problem, any one of the autocorrelation estimators will do. For instance, given the equations that we present, a statistical necessity is that the within-study estimates will shrink toward a common mean, depending solely on how much sampling error is present, no matter which autocorrelation estimator is used. This need not be demonstrated with, say, a simulation, though we must be clear that the Bayesian estimates in this article clarify only the precision of the estimate, not its bias. Second, the focus of the present study is not on reducing bias in the autocorrelation, but on demonstrating how the Bayesian approach is capable of dealing with sampling error in the measurement of autocorrelations in a way that the usual frequentist statistics cannot do.

Assume that one has  $k$  cases, each consisting of  $t_j$  data points, each case producing an observed autocorrelation  $r_j$  ( $j = 1 \dots k$ ). Then, the EB estimate  $\hat{\rho}_j^*$  of the  $j$ th autocorrelation can be defined as a weighted sum of information from two sources:

$$\hat{\rho}_j^* = \lambda_j r_j + (1 - \lambda_j) \hat{\rho}_., \quad (2)$$

in which equation  $\hat{\rho}_.$  is the random-effects meta-analytic average autocorrelation from the full set of SCDs, and  $\lambda_j$  is the reliability of the estimate of a case's autocorrelation  $r_j$ :

$$\lambda_j = \frac{Var(\hat{\rho}_.)}{Var(r_j)} = \frac{\tau}{\tau + v_j}. \quad (3)$$

Here,  $\tau$  is the variance of the estimated autocorrelations across studies, and  $v_j$  is the error variance of the sample statistic defined in Eq. 1.

This definition of reliability is essentially the same as that used in classical test theory: the ratio of true score variance to observed score variance. Hence, the EB estimate of autocorrelation in Eq. 2 is a weighted composite of the observed autocorrelation from the given SCD of interest and the average autocorrelation from the population from which the set of studies for the given SCD is drawn, where the weight is the reliability of the respective autocorrelations. In the context of SCDs, we would typically expect the reliability of the autocorrelation from a given SCD to be quite low due to sampling error. We would also expect that the reliability of the average of the set of autocorrelations would be higher—sometimes considerably higher—with the reliability of that average increasing monotonically with the number of the SCDs and time points that contribute to the average. Hence, the observed estimate from any given SCD will move closer to the average autocorrelation, reflecting the greater reliability of the latter. Consequently, these EB estimates are often referred to as *shrunken estimates*, because each EB estimate shrinks toward the mean. Unlike the meta-analytic average, the EB estimates can be unique to each SCD. That is, where the meta-analytic approach might substitute a single number (the average) for each of  $k$  autocorrelations, the EB approach frequently yields  $k$  distinct autocorrelations, and when it does not do so, this is because a single population parameter is the most plausible estimate for the underlying autocorrelation.

A subsidiary issue is that the asymptotic variance in Eq. 1 may be biased, which could affect the degree of shrinkage. For instance, if the variance in Eq. 1 underestimates the true variance, the reliability of the sample autocorrelation in Eq. 3 may be overestimated, and Eq. 2 may then yield an estimate that is not shrunken as much as would be the case if a more accurate variance estimate were available. If the variance is overestimated, the opposite will happen. Unfortunately, no widely accepted alternative estimate of the variance currently exists. How much any such shrinkage may be affected will be a useful topic for further investigation.

## Computations

To obtain the EB estimates, one first needs to estimate the autocorrelation for each SCD. In the examples, we will use a standard estimate of lag-1 autocorrelation:

$$r_j = \frac{\sum_{t=1}^{n-1} y_t y_{t+1}}{\sum_{t=1}^n y_t^2}, \quad (4)$$

where  $y_t$  is the residual of the observation at time  $t_j$ , and  $y_{t+1}$  is the residual at time  $t_{j+1}$  (Huitema & McKean, 1994). The remaining information to compute the EB estimates of these autocorrelations is obtained by computing a meta-analysis on the autocorrelations (Shadish & Haddock, 2009). The random-effects average effect size  $\hat{\rho}_.$  is estimated as

$$\hat{\rho}_. = \frac{\sum_{j=1}^k w_j r_j}{\sum_{j=1}^k w_j}, \quad (5)$$

where  $w_j = 1/v_j^*$  and  $v_j^* = \tau + v_j$ . The last variable was already computed as Eq. 1, and

$$\tau = \max \begin{cases} [Q - (k - 1)]/c, \\ 0 \end{cases} \quad (6)$$

where

$$c = \sum_{j=1}^k w_j - \left[ \sum_{j=1}^k w_j^2 \middle/ \sum_{j=1}^k w_j \right] \quad (7)$$

and

$$Q = \sum_j w_j (\hat{\rho}_j - \hat{\rho}_.)^2. \quad (8)$$

The results are inserted into Eqs. 2 and 3 to obtain an EB estimate for each SCD (SPSS and R scripts to do these computations are available as [supplemental materials](#)).

## Examples

We illustrate these methods with two examples that illustrate conditions under which EB estimates work with differential effectiveness. Schutte, Malouff, and Brown (2008) used a multiple-baseline design to study the efficacy of an emotion-focused therapy in 13 adults suffering

from prolonged fatigue. The number of time points ranged from six to 15 over cases. The authors included all of the raw data in a table in the publication. We created residuals for each SCD from a regression of the outcome on time, treatment, and an interaction term. The rationale for these predictors was that a primary regression analysis of such data would include these terms as predictors, and those terms would account for some portion of the autocorrelation. The question at issue would then be the size of any remaining (residual) autocorrelation once those terms were modeled.

Because the number of time points in each SCD in the Schutte et al. (2008) data was small, the autocorrelation from any given SCD was likely to be measured with great sampling error. However, considerably more information about the autocorrelation was available from the set of 13 SCDs in the Schutte et al. data, so we would expect that the EB estimates would shrink toward the overall meta-analytic mean from the 13 studies. This was indeed the case. The raw autocorrelations computed by Eq. 4 ranged from  $r_j = -.671$  to  $+.354$  (see Table 1). A meta-analytic homogeneity test indicated that these autocorrelations differed by no more than we would expect by chance ( $Q = 9.37$ ,  $df = 12$ ,  $p = .67$ ), and  $\hat{\tau} = 0$ ,

**Table 1** Autocorrelations using different estimators

Schutte et al. (2008)	$t_j$	$r_j$	$r_{EB}$	$r_{HLMEB}$
Case 1	12	-.398	-.172	-.172
Case 2	15	.354	-.172	-.172
Case 3	9	-.083	-.172	-.172
Case 4	9	-.149	-.172	-.172
Case 5	15	-.064	-.172	-.172
Case 6	12	-.556	-.172	-.172
Case 7	12	-.343	-.172	-.172
Case 8	10	.204	-.172	-.172
Case 9	9	-.124	-.172	-.172
Case 10	9	-.331	-.172	-.172
Case 11	15	-.165	-.172	-.172
Case 12	6	-.671	-.172	-.172
Case 13	12	-.101	-.172	-.172
Summary statistics for $r$ : $\hat{\rho}_. = -.1718$ , $\tau = .000$ , $c = 108.823$ , $Q = 9.37$				
Dyer et al. (1984)	$t_j$	$r_j$	$r_{EB}$	$r_{HLMEB}$
Case 1	47	-.184	-.152	-.151
Case 2	47	-.202	-.165	-.164
Case 3	47	.380	.284	.281
Case 4	47	-.216	-.176	-.175
Summary statistics for $r$ : $\hat{\rho}_. = -.0525$ , $\tau = .0682$ , $c = 141.622$ , $Q = 12.65 *$				

$t_j$  is the number of time points,  $r_j$  is the usual autocorrelation,  $r_{EB}$  is the empirical Bayes estimate of  $r_j$ , and  $r_{HLMEB}$  is the HLM empirical Bayes estimate of  $r_j$ . \*  $p < .05$

so that Eq. 2 reduced to the meta-analytic random-effects mean. Consequently, all estimates shrank to that mean,  $r_{EB} = -.172$ . The large variability and extreme values of the autocorrelation in the Schutte et al. data had been entirely due to sampling error.

The second example is from Dyer, Schwartz, and Luce (1984), who trained severely handicapped students to engage in age-appropriate and functional activities. Using a multiple-baseline design, they implemented training in four homes sequentially, with one baseline, one treatment, and one maintenance phase in each home. Ignoring the maintenance phase data, we digitized the data from these four graphs using procedures outlined elsewhere that have extremely high reliability and validity (Shadish et al., 2009), and we again created residuals for each SCD from a regression of the outcome (average score per student) on time, treatment, and an interaction term. In the Dyer et al. data, the number of data points in each SCD was relatively large, at  $t_j = 47$ , so the information in each SCD about the autocorrelation was considerably more reliable than was the case for the Schutte et al. (2008) data, and the EB estimates showed less shrinkage. Otherwise, the results are similar to those for the Schutte et al. data. The raw autocorrelations for the Dyer et al. data computed by Eq. 4 ranged from  $r_j = -.216$  to  $+.380$ , and they were heterogeneous ( $\hat{\tau} = .0682$ ,  $Q = 12.65*$ ); the EB estimates of Eq. 4 ranged from  $r_{EB} = -.176$  to  $+.284$ , or about 23 % shrinkage. We note one unusual aspect of the Dyer et al. data that might affect inferences: Three of the autocorrelation estimates were very similar, but the fourth was quite different, and may be considered an outlier. House 1 in that study had the discrepant autocorrelation. Inspection of the graph for that case suggests the possibility of a ceiling effect that was not present in the other three cases, and this could cause higher autocorrelation.

EB estimates can also be obtained from multilevel models using programs that are more convenient for some researchers than SPSS script. We did so using the HLM computer program (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004). Using the *v-known* option in HLM, the only inputs needed are the autocorrelations and their conditional variances; HLM computes all of the other information internally (i.e., Eqs. 2, 3, 5, 6, and 7). For the Schutte et al. (2008) data, the EB estimates ( $r_{HLMEB}$ ) are identical to the  $r_{EB}$  estimates. The variance estimate computed directly by HLM ( $\hat{\tau} = .0003$ ) is near zero, and probably would be exactly zero if one reduced the default tolerance of HLM for convergence (the EM algorithm that it uses goes to zero values slowly). So the HLM EB estimates are shrunk completely toward the mean. Similarly, for the Dyer et al. (1984) data, the HLM EB estimates are again nearly identical to the  $r_{EB}$  estimates. Finally, we also have an R script for these

computations, although R has a much steeper learning curve than do the other two programs. SPSS and R syntax for this analysis are available as [supplemental materials](#).

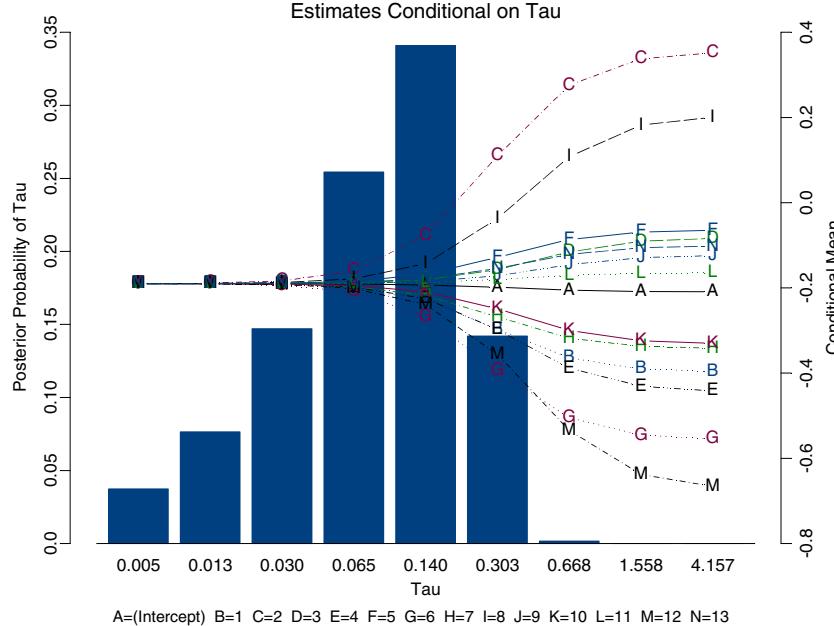
## Fully Bayesian estimation

Fully Bayesian methods are potentially very useful for single-case design research, because of the small number of cases and time points in most studies. The EB estimates first find an estimate of the true variability in the autocorrelations among cases ( $\tau$ ), and then use that estimate in Eq. 3 as if it were completely accurate when finding the shrunken estimates of the autocorrelations. In fact, however,  $\tau$  is itself estimated, often with considerable uncertainty, especially in small samples; fully Bayesian methods take this uncertainty into account. The result is that one does not get just one shrunken estimate of the autocorrelation, as in the EB approach, but a posterior distribution of shrunken estimates that depends on the estimated distribution of  $\tau$ . We used these methods on the two example data sets. The prior for the average autocorrelation is uninformative, and the prior for  $\tau$  is a Pareto density with form  $f(\tau) = k/(k + \tau)^2$ , where  $k$  is the square root of the harmonic mean of the variances of the autocorrelations. This is relatively uninformative for  $\tau$ , but discounts very large values of  $\tau$ .

The resulting interpretation is considerably more nuanced and informative, although the general tenor of the results is unchanged. For example, with the Schutte et al. (2008) data, the weighted average shrunken estimates ranged from  $-.25$  to  $-.10$  instead of being fully shrunken, as were the EB estimates. The posterior mean of the autocorrelation is  $r_j = -.19$  ( $SD = .1005$ ,  $p = .09$ ), and the posterior mean of  $\tau$  is  $.11$  ( $SD = .09$ ). Those with the desire to use fully Bayesian methods can do so through software such as the program that we used here, hblm in SPlus (DuMouchel, 1994, 1995), or WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). The SPlus syntax for this analysis is available as [supplemental materials](#).

The nature of the fully Bayesian analysis is most easily explained in terms of a picture. Figure 1 contains a trace plot of the results of the analysis of the Schutte et al. (2008) data set produced by the hblm software (DuMouchel, 1994, 1995). The bars illustrate the distribution of probabilities about possible values of the residual standard deviation  $\tau$ . Most plausible values (those with large bars) of  $\tau$  are small, corresponding to  $\tau$  variances less than  $.09$ , though they seem to indicate that  $\tau$  is unlikely to be zero. The lines with letters correspond to either the overall mean (indicated by the letter A) or individual autocorrelations from the Schutte et al. data

**Fig. 1** Trace plot for the Schutte et al. (2008) data, showing the posterior distribution of  $\tau$  (on the left vertical axis) and autocorrelation estimates conditional on each value of  $\tau$  (on the right vertical axis). The line labeled A is the overall average, while B through N represent estimates for each individual autocorrelation



set (letters B through N). At each of nine values of  $\tau$ , the letters show what the shrunken estimates of the autocorrelation would be if that value of  $\tau$  were the true value (with the height of the bar indicating the probability that  $\tau$  is the true value). Small values of  $\tau$  correspond to maximal shrinkage, so the lines converge on the left side of the plot. As  $\tau$  gets larger, the lines diverge, and on the far right approach the observed effect sizes (corresponding to no shrinkage). An empirical Bayes method picks the most likely value of  $\tau$  and estimates shrunken values of the autocorrelation at that value of  $\tau$ . A fully Bayesian method averages the autocorrelations across the possible values of  $\tau$ , weighting according to how likely each value of  $\tau$  is to be correct. (Technically, because  $\tau$  is continuously distributed, this is an integral, but considering nine discrete values and summing, as was done here, produces a very accurate estimate.)

One additional consequence of not assuming  $\tau$  to be known (or, equivalently, estimated with complete precision) is that the confidence intervals for the shrunken estimates and parameter estimates are (appropriately) wider because of the additional uncertainty. Typically, one wants the smallest standard errors and narrowest confidence intervals possible, but only if these are honest (e.g., in frequentist terms, a 95 % confidence interval really covers the population value 95 % of the time, not 80 % of the time). Fully Bayesian intervals are honest in this sense.

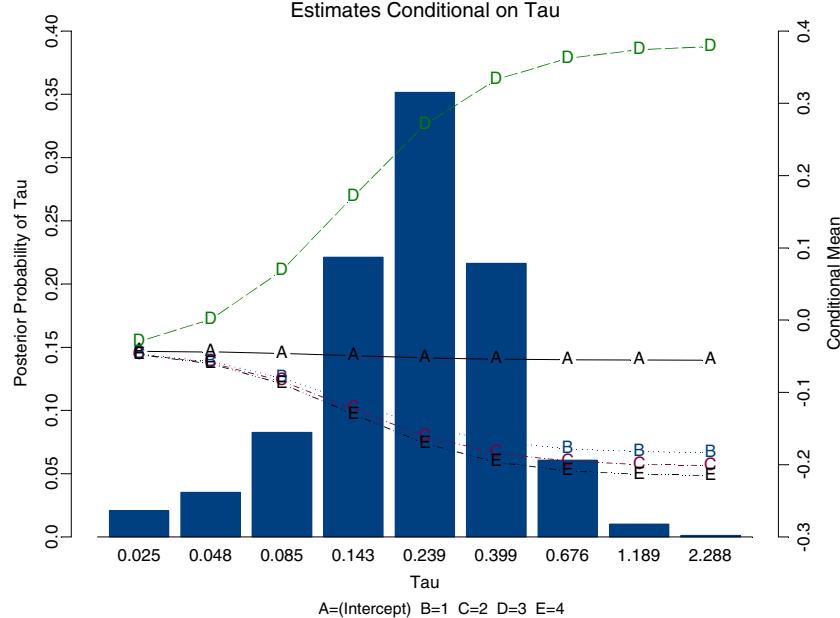
The trace plot for the Dyer et al. data (see Fig. 2) reveals a problem that was apparent previously in Table 1. Three autocorrelations have estimates that are very similar, so shrinking them toward a common value for the purpose of strengthening the evidence about each is not problematic. The fourth value, however, has an estimate that is far from the other three, as can be seen by the divergence of the lines

as one goes from left to right in the figure. It does not seem reasonable to combine that value with the other three, as it seems clearly to be an outlier. The statistical evidence is also consistent with this judgment, as is the previous observation that this case suffered from a ceiling effect that could have caused the discrepant autocorrelation. In some cases that are less extreme than this one, a  $t$  distribution might be more reasonable than a normal distribution, as this allows heavier tails than does the normal distribution. Alternatively, one might treat this as a case that does not meet the exchangeability criterion and exclude it from the Bayesian analyses.

## Discussion

Bayesian estimates have one main advantage as compared to past estimators of the autocorrelation in SCDs: They reduce the role of sampling error in the estimates, particularly when the number of cases is low, or the number of data points within a case is small, or both. Hence, Bayesian estimates account better for the variability in the distribution of autocorrelations. When faced with the choice of using an individual sample autocorrelation or the Bayesian estimate in a statistic that requires adjustment for autocorrelation, the latter is more likely than the individual sample autocorrelations to represent the variability of the population autocorrelations correctly. For example, Hedges et al. (2012) developed a  $d$  statistic for SCDs, and knowledge of the autocorrelation is necessary to compute both the denominator of that statistic and its conditional variance. In such cases, a Bayesian estimate of the autocorrelation will, on average, lessen the role of sampling error in autocorrelations. Even for descriptive studies of autocorrelations (e.g.,

**Fig. 2** Trace plot for Dyer et al. (1984) data, showing the posterior distribution of  $\tau$  (on the left vertical axis) and autocorrelation estimates conditional on each value of  $\tau$  (on the right vertical axis). The line labeled A is the overall average, while B through E represent estimates for each individual autocorrelation



Parker et al., 2005), providing Bayesian estimates of the autocorrelation instead of, or in addition to, the usual sample estimates will separate issues of parameter estimation from those of sampling error.

We can gain some insight about how autocorrelations of different levels affect bias from a simulation by Hedges et al. (2012) of bias in the estimates of their SCD  $d$  statistic and its variance. The simulation involved five parameters, only three of which need concern us here. The number of cases,  $m$ , was varied from four to 12 to capture a range of values observed in practice. The number of observations per period,  $n$ —assumed to be equal in the baseline and treatment periods—was varied from four to 12. Finally, they varied the autocorrelation over all but the most extreme possible values. For each combination of the parameters, they simulated 8,000 iterations of the model. Briefly, the results confirmed that the bias of  $G$  ( $d$  corrected by a small sample-bias correction) remains small, except in the case of very large (and probably unrealistic) negative autocorrelations (e.g.,  $-.9$ ). The variance of  $G$  was estimated somewhat more poorly, especially when  $n = m = 4$ . As the number of time points increased, bias in the variance became minimal when the autocorrelation was less than  $|\pm .25|$  and  $m = n = 12$ . The Bayesian estimators of the autocorrelation in the present article are low enough to suggest that the autocorrelation may not much bias results for this  $d$  statistic itself, but for its variance, autocorrelation could prove to be a problem at lower sample sizes. This clearly needs more investigation.

Several variations and extensions on these methods are possible. First, the variance estimate for the autocorrelation (as for any other correlation) depends on the value of the correlation itself. In substituting the observed autocorrelation, we typically underestimate the variance because we

have not yet shrunk the estimates; thus, extreme values of correlations are not shrunk enough. One could have a two-step procedure, in which the variances are reestimated after the initial shrinkage, and then the estimates are recomputed. In other words, variances from shrunken estimates are used in the analysis, but with the original autocorrelations. This would make the standard errors of the autocorrelations more accurate. Another possibility is to use a Fisher's  $Z$  transformation of the correlations, which has a variance estimate that is not a function of the correlation. (See Hafdahl, 2009, 2010, for more on the use of Fisher's  $Z$  in meta-analysis.)

The two examples that we have used in this article provide some insight into the autocorrelation problem in SCDs. High variability in values of observed autocorrelations may easily be taken as indicating a problem when it should not. For both Schutte et al. (2008) and Dyer et al. (1984), the estimates of the average autocorrelations are close to zero, and the extreme values observed in the individual study autocorrelations are generally smaller with the shrunken estimates. Though this is only evidence from two studies, and so should not be generalized with any confidence, they do show that it is very easy to think that autocorrelation is a bigger problem with SCDs than it may be. How big the problem genuinely may be needs empirical investigation.

SCD researchers who are concerned about autocorrelation would be well-advised to compute meta-analytic statistics on the autocorrelations from the set of studies that they are using. This typically might include the random-effects average autocorrelation, including whether that average is significantly different from zero, and also the significance of a homogeneity test to tell whether the variability in autocorrelations over studies is greater than would be expected by

chance. When neither significance test is rejected, one could proceed on the assumption that autocorrelation may not be a problem in that set of studies. However, since the power of the homogeneity test can be low in small samples, a better approach might be to always compute either the empirical or fully Bayesian estimates.

This study has several limitations. First, in the SPSS script for the empirical Bayes estimates, we used a method-of-moments estimate of the between-studies variance ( $\tau^2$ ) of  $\hat{\rho}$ , but a full or restricted maximum likelihood estimate might be more accurate. We doubt that this would make much difference to the general pattern of results that we report, especially since the HLM EB estimates obtained using an EM algorithm were nearly identical to the method-of-moments results. Second, the estimate of the between-studies variance ( $\tau^2$ ) of  $\hat{\rho}$  has low precision in small sample sizes (Shadish & Haddock, 2009), so its precision will likely be in question in many applications to SCDs. This is partly why the fully Bayesian estimates can be so informative, given that they look at autocorrelation over the probable range of  $\tau$ . Third, when homogeneity of autocorrelations was rejected, as it was for the Dyer et al. (1984) study, we did not investigate possible sources of that heterogeneity. We did note that House 1 in that study had a discrepant autocorrelation, and that inspection of the graph for that case suggested the possibility of a ceiling effect that was not present in the other three cases, which could cause a higher autocorrelation. Fourth, information about the autocorrelations from two studies is inherently limited. A larger meta-analytic study of autocorrelations in the SCD literature would help clarify the conditions under which autocorrelation might or might not be a problem.

**Author note** This research was supported in part by Grant Nos. R305D100046 and R305D100033 from the Institute for Educational Sciences, U.S. Department of Education, and by a grant from the University of California Office of the President to the University of California Educational Evaluation Consortium. The opinions expressed are those of the authors and do not represent views of the University of California, the Institute for Educational Sciences, or the U.S. Department of Education.

## References

Anderson, T. W. (1971). *The statistical analysis of time series*. New York, NY: Wiley.

DuMouchel, W. (1994). *Hierarchical Bayes linear models for meta-analysis*. NISS Technical Report Number 27. Research Triangle Park, NC: National Institute of Statistical Sciences.

DuMouchel, W. (1995). *Documentation for Hierarchical Bayes Linear Model Programs*. Downloaded from [ftp://ftp.research.att.com/dist/bayes-meta/hblm\\_doc.ps](ftp://ftp.research.att.com/dist/bayes-meta/hblm_doc.ps) February 18, 2011.

Dyer, K., Schwartz, I. S., & Luce, S. C. (1984). A supervision program for increasing functional activities for severely handicapped students in a residential setting. *Journal of Applied Behavior Analysis*, 17, 249–259. doi:10.1901/jaba.1984.17-249

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care*, 49, 761–768. doi:10.1097/MLR.0b013e31825f6619

Hafdahl, A. R. (2009). Improved Fisher  $z$  estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, 62, 233–261. doi:10.1348/000711008X281633

Hafdahl, A. R. (2010). Random-effects meta-analysis of correlations: Monte Carlo evaluation of mean estimators. *British Journal of Mathematical and Statistical Psychology*, 63, 227–254. doi:10.1348/000711009X431914

Hedges, L. G., Pustejovsky, J., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods*, 3, 224–239. doi:10.1002/jrsm.1052

Huitema, B. E., & McKean, J. W. (1994). Two biased-reduced autocorrelation estimators:  $r_{F1}$  and  $r_{F2}$ . *Perceptual and Motor Skills*, 78, 323–330. doi:10.2466/pms.1994.78.1.323

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, 60, 38–58. doi:10.1177/00131640021970358

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis*, 11, 277–283. doi:10.1901/jaba.1978.11-277

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15, 124–144. doi:10.1037/a0017736

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337. doi:10.1023/A:1008929526011

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, 49, 301–321. doi:j.jsp.2011.03.004/j.jsp.2011.03.004

Manolov, R., & Solanas, A. (2008). Comparing  $N = 1$  effect size indices in presence of autocorrelation. *Behavior Modification*, 32, 860–875. doi:10.1177/0145445508318866

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double Bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, 5, 87–101. doi:10.1037/1082-989X.5.1.87

Parker, R. I., Brossart, D. R., Vannest, K. J., Long, J. R., De-Alba, R. G., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review*, 205, 116–132.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011a). Effect size in single case research: A review of nine nonoverlap techniques. *Behavior Modification*, 35, 303–322. doi:10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011b). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, 42, 284–299. doi:j.jsp.2011.03.004/j.beth.2010.08.006

Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Schutte, N. S., Malouff, J. M., & Brown, R. F. (2008). Efficacy of an emotion-focused treatment for prolonged fatigue. *Behavior Modification*, 32, 699–713. doi:10.1177/0145445508317133

Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to

extract data from image files: Verification of reliability and validity. *Behavior Research Methods*, 41, 177–183. doi:[10.3758/BRM.41.1.177](https://doi.org/10.3758/BRM.41.1.177)

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 257–277). New York, NY: Russell Sage Foundation.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. In G. Julnes & D. J. Rog (Eds.), *Informing federal policies on evaluation method: Building the evidence base for method choice in government sponsored evaluation* (pp. 95–109). San Francisco, CA: Jossey-Bass.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 3, 188–196. doi:[10.1080/17489530802581603](https://doi.org/10.1080/17489530802581603)

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, 43, 971–980. doi:[10.3758/s13428-011-0111-y](https://doi.org/10.3758/s13428-011-0111-y)